Continual Learning for Long-Tailed Recognition: Bridging the Gap in Theory and Practice

Mahdiyar Molahasani, Ali Etemad & Michael Greenspan
Smith School of Engineering and Ingenuity Labs Research Institute
Queen's University
Kingston, Canada
{m.molahasani,ali.etemad,michael.greenspan}@queensu.ca

Abstract

The Long-Tailed Recognition (LTR) problem arises in imbalanced datasets. This paper bridges the theory-practice gap in this context, providing mathematical insights into the training dynamics of LTR scenarios by proposing a theorem stating that, under strong convexity, the learner's weights trained on the full dataset are bounded by those trained only on the Head. We extend this theorem for multiple subsets and introduce a novel perspective of using Continual Learning (CL) for LTR. We sequentially learn the Head and Tail by updating the learner's weights without forgetting the Head using CL methods. We prove that CL reduces loss compared to fine-tuning on the Tail. Our experiments on MNIST-LT and standard LTR benchmarks (CIFAR100-LT, CIFAR10-LT, and ImageNet-LT) validate our theory and demonstrate the effectiveness of CL solutions. We also show the efficacy of CL on real-world data, specifically the Caltech256 dataset, outperforming state-of-the-art classifiers. Our work unifies LTR and CL and paves the way for leveraging advances in CL to tackle the LTR challenge effectively.

1 Introduction

Data in real-world scenarios often exhibits long-tailed distributions [1, 2, 3, 4], where the number of samples in some classes (Head set) is significantly larger than in other classes (Tail set). This imbalance of data distribution presents challenges for the optimization and generalization of deep learning models, described as the Long-Tailed Recognition (LTR) problem, which is defined as training a model on highly imbalanced data while aiming for high performance on a balanced test set [3]. In this context, optimizing the training process becomes a non-trivial task. The Head classes, due to their sample dominance, disproportionately influence the loss function and the gradient updates. This often results in a model that performs well on the Head but poorly on the Tail [5]. Numerous studies have addressed the issue of class imbalance by various methods: over-sampling Tail classes [6, 7, 8], employing feature extractors trained on the Head set for transfer learning [9, 10, 11, 12], regularizing loss or gradients [13, 14, 15], and recently, weight balancing to maintain uniform per-class weight norms [5]. As machine learning models continue to grow in size and complexity, understanding the mathematical properties that govern their training dynamics becomes critical.

This paper proposes a novel perspective that unifies the problems of Continual Learning (CL) and LTR, which facilitates the application of CL solution approaches directly to LTR problems. CL aims to minimize forgetting when deep learning models are adapting themselves to new tasks/distributions. Consequently, we contribute to unifying these two areas by proving a theorem stating that under strong convexity, the learner's weights trained on the full dataset are confined within an upper bound relative to those trained solely on the Head. This bound is proportional to the dataset's imbalance factor and inversely proportional to the loss function's strong convexity param-

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

eter. We extend this theorem to arbitrary partitions with varying class sizes, proving that weights from training on these subsets also lie within bounded neighborhoods of those trained on the largest subset. Subsequently, we propose using Continual Learning (CL) methods to sequentially learn the Head and Tail without forgetting the former. We introduce a further theorem, demonstrating that CL yields a lower loss when compared to strictly fine-tuning on the Tail. Our theory is validated on five datasets: MNIST-LT, CIFAR100-LT, CIFAR10-LT, ImageNet-LT, and Caltech256. Experiments show that CL methods achieve effective performances compared to baselines and SOTA LTR models, particularly on the naturally imbalanced Caltech256 dataset.

Our contributions in this paper can be summarized as follows: (**1**) We propose a mathematical insight into the optimization dynamics in the LTR scenario by establishing an upper bound on the distance between weights obtained when trained on the full dataset and the Head. Furthermore, we extend this theorem to apply to any number of partitions with varying class sizes. (**2**) Using this bound as a basis, we introduce a new perspective on using CL solutions for the LTR problem supported by another theorem that proves the effectiveness of CL in reducing the loss when focusing on Tail classes. (**3**) We substantiate our method through comprehensive experiments that demonstrate the effectiveness of CL techniques in addressing LTR. Our results indicate significant performance gains in long-tailed scenarios when using standard CL approaches.

2 Method

2.1 Overview

Assume an LTR problem with a learner, denoted as θ (initialized with θ_i), trained on an imbalanced dataset $\mathcal{D}$, as shown in Fig. 1. The gradients are dominated by the larger Head set, leading the learner's parameters to converge to θ^*. We propose a theorem stating that under a strongly convex loss function, θ^* lies within a bounded radius r of θ^*_H, the weights when trained solely on the Head set $\mathcal{D}_H$. The radius r is proportional to the loss function's strong convexity and inversely proportional to the imbalance factor. To address this, we reformulate LTR as two sequential tasks: learning the Head and Tail separately. However, sequential learning introduces catastrophic forgetting. Ideal weights θ^*_{HT} lie in the intersection of ψ_H (Where model performs well on Head) and ψ_T (Where model performs well on Tail), denoted by ψ_{HT}. To avoid forgetting the Head while learning the Tail, we employ Continual Learning (CL) techniques. By using CL, the model learns the Tail set without compromising its performance on the Head, ultimately performing well on both sets and ending up in ψ_{HT}.

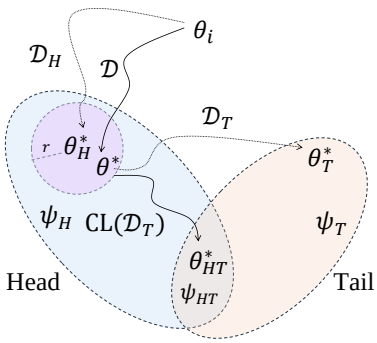


Figure 1: An overview of learning under the LTR scenario and our proposed algorithm is presented. Detailed description provided in text.

2.2 Problem Formulation

LTR addresses learning from imbalanced data, segmented into Head set ($\mathcal{D}_H$) with more samples and Tail set ($\mathcal{D}_T$) with fewer. Imbalance factor (IF) quantifies the severity of this issue in a dataset and is defined as $\frac{|\mathcal{D}_{c^{\max}}|}{|\mathcal{D}_{c^{\min}}|}$, where $c^{\max} = \arg\max |\mathcal{D}_c|$, and $c^{\min} = \arg\min |\mathcal{D}_c|$, such that $\mathcal{D}_{c^{\max}} \in \mathcal{D}_H$ and $\mathcal{D}_{c^{\min}} \in \mathcal{D}_T$.

Definition 2.1. *A dataset is deemed* long-tailed *when* $|\mathcal{D}_{c^{\max}}| \gg |\mathcal{D}_{c^{\min}}|$ *or, in other words, IF* $\gg 1$. *When a model is trained on such a dataset and its performance is assessed on a uniformly distributed test set (i.e.* $|\mathcal{D}_c| = k$ *for each class* $\mathcal{D}_c$ *within the test set), the problem is referred to as* Long-Tailed Recognition.

2.3 Training on Long-tailed Distribution

In this section, we derive the conditions in which CL can be applied to a long-tailed scenario. We assume that all head classes are of size $|\mathcal{D}_H|$, and all tail classes are of size $|\mathcal{D}_T|$, with $|\mathcal{D}_H| >> |\mathcal{D}_T|$. The training process in the LTR setup is analyzed using the following Theorem.

Theorem 2.2. *Assume that a logistic regression model with parameters θ is trained using regularized cross-entropy loss in an LTR setting. Then, $\|\theta^* - \theta_H^*\|^2 \leq \frac{4\delta}{\mu_H + \mu}$, where θ^* represents the parameter vector obtained after training, θ_H^* denotes the parameter vector when the model is trained solely on the Head set, δ is the maximum difference between the loss of the learner using the entire dataset or the Head set for any value of θ, and μ_H and μ are the strong convexity parameters of the loss computed on either the Head set or the entire dataset.*

Proof Sketch. (Formal proof in Appendix B.1) Initially, we decompose the total loss into two components: one for the Head and another for the Tail. We establish that in a LTR setting, the total loss asymptotically converges to the Head loss. Subsequently, leveraging the strong convexity property of the loss function, we prove that as these two loss components converge, their respective minimizers also converge.

The analysis in Appendix B.2 confirms our findings on the upper bound, even when assuming strict convexity instead of strong convexity. Theorem 2.2 is based on a single Head and Tail in the dataset, which is often not true in real-world data. To generalize, we allow the Head and Tail sets to follow a long-tailed distribution and partition them into multiple Head and Tail subsets. We continue this partitioning until the imbalance factor IF_{D^i} for each subset D^i is not significantly greater than 1. Theorem 2.3 extends Theorem 2.2 to cover this more complex scenario.

Theorem 2.3. *Let a logistic regression model with parameters θ be trained using regularized cross-entropy loss in an LTR setting, and let dataset $\mathcal{D}$ be divided into n partitions. Further, let a subset of $m < n$ partitions be $\bigcup_{i=1}^{m} \mathcal{D}_i \subseteq \mathcal{D}$, with the largest partition being $\mathcal{D}_a$, i.e. $a = \arg\max_i |\mathcal{D}_i|, i \in [1, m]$. Then, the weights $\theta^*_{\bigcup \mathcal{D}_i}$ obtained from training the model on $\bigcup_{i=1}^{m} \mathcal{D}_i$ will always be in a bounded neighborhood of the weights $\theta^*_{\mathcal{D}_a}$ obtained from training on the largest subset $\mathcal{D}_a$.*

Proof Sketch. (Appendix B.5) We partition the dataset and focus on the two largest subsets. Using Theorem 2.2, we derive a bound for weight differences between them. Iteratively, we aggregate the largest subset with the next largest, applying Theorem 2.2 each time to calculate an upper bound on weight differences for training on all subsets versus the largest aggregated subset.

2.4 CL for LTR

In order to prove the effectiveness of employing CL methods for addressing LTR problems, the following theorem is proposed.

Theorem 2.4. *Consider a logistic regression model with parameters θ trained using regularized cross-entropy loss ($\mathcal{L}$) in an LTR setting, converging to θ^i. Then, $\mathcal{L}(\mathcal{D}, \theta_{EWC}^{i+1}) < \mathcal{L}(\mathcal{D}, \theta_{\mathcal{L}}^{i+1})$, where θ_{EWC}^{i+1} and $\theta_{\mathcal{L}}^{i+1}$ denote the weights of the model after a single update using Elastic Weight Consolidation (EWC) [16] loss and regularized cross-entropy loss, respectively.*

Proof Sketch. (See Appendix B.6) Using Taylor expansion, we approximate losses for weights updated via EWC and regularized cross-entropy. We prove EWC's regularization term constrains weight updates more effectively. Due to the strong convexity and the Fisher information matrix's positive nature, EWC-updated weights yield a strictly lower loss.

This theorem validates that even a fundamental CL method like EWC can enhance LTR. Its mathematical simplicity makes it suitable for Theorem 2.4, suggesting that advanced CL methods will likely offer further benefits, as confirmed by subsequent experimental results.

3 Experimental Results and Discussions

Upper bound. To verify the upper bound in Eq. 8, we calculate the estimated upper bound for each γ and μ using Eq. 14 in Appendix B.3. It is important to note that this upper bound is tighter compared to Eq. 8 . We compare the upper bound with the actual distance in Fig. 2 and show that for all *IF* and μ values, the measured distance is lower than the theoretical upper bound.

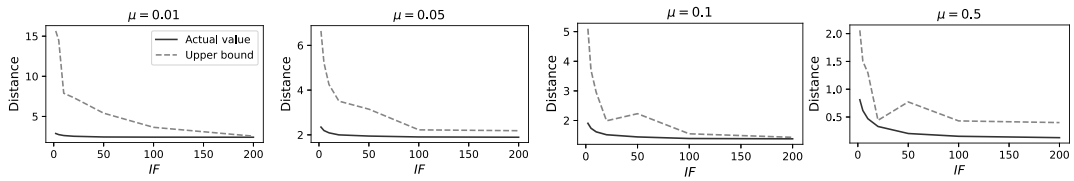


Figure 2: The actual distance between θ^* and θ^*_H in different *IF* and μ compared with the calculated upper bound.

Table 1: LTR benchmarks for CIFAR100-LT.

Model	IF 100	50	10
Baseline [17]	38.32	43.85	55.71
Baseline + CB [17]	39.60	45.32	57.99
LTR methods			
Focal loss [18]	38.41	44.32	55.78
Focal+CB [17]	39.60	45.17	57.99
τ-norm [19]	47.73	52.53	63.81
LDAM-DRW [20]	42.04	46.62	58.71
BBN [21]	42.56	47.02	59.12
LogitAjust [22]	42.01	47.03	57.74
LDAM+SSP [23]	43.43	47.11	58.91
De-confound [24]	44.10	50.30	59.61
SSD [25]	46.00	50.50	62.30
DiVE [26]	45.35	51.13	62.00
DRO-LT [27]	47.31	57.57	63.41
PaCo [28]	52.00	56.00	64.20
WD [5]	46.01	52.71	66.03
WD & Max [5]	53.35	57.71	68.67
CL methods			
LwF [29]	45.05	49.33	58.71
EWC [16]	44.35	50.28	58.84
Modified EWC	45.93	50.98	60.67
GPM [30]	47.93	53.20	63.31
SGP [31]	50.04	55.91	66.13

Table 2: LTR benchmarks for CIFAR10-LT.

Model	IF 100	50
Baseline [17]	69.8	75.2
Baseline + CB [17]	74.7	79.3
LTR methods		
Mixup [32]	73.1	77.8
Focal loss[18]	70.4	75.3
PG Re-sampling [33]	67.1	75.0
3LSSL [34]	85.2	88.2
Focal+CB[17]	74.6	79.3
LDAM-DRW[20]	77.0	79.3
BBN [35]	79.8	82.2
Manifold mixup [17]	73.0	78.1
CBA-LDAM [17]	80.3	82.2
ELF (LDAM)+DRW [17]	78.1	82.4
De-confound [24]	80.6	83.6
Hybrid-SC [36]	81.4	85.4
MiSLAS [37]	82.1	85.7
BCL [38]	84.3	87.2
CL methods		
LwF [29]	76.3	78.6
EWC [16]	75.1	80.1
Modified EWC	77.8	81.3
GPM [30]	81.2	84.8
SGP [31]	83.0	85.5

Table 3: LTR benchmarks for ImageNet-LT.

Model	Top-1 accuracy
Baseline [17]	44.4
Baseline + CB [17]	33.2
LTR methods	
KD [39]	35.8
Focal [18]	30.5
SR Re-sampling [40]	46.8
OLTR [41]	35.6
cRT [19]	49.6
τ-norm [19]	49.4
LFME [42]	37.5
De-confound [24]	51.8
Seasaw Loss [43]	50.4
DiVE [26]	53.1
DRO-LT [27]	53.5
DisAlign [44]	52.9
WD [5]	48.6
WD+Max [5]	53.9
CL methods	
LwF [29]	47.6
EWC [16]	48.9
Modified EWC [45]	49.1
GPM [30]	50.6
SGP [31]	52.0

LTR datasets To validate the efficacy of CL in LTR, we apply five CL strategies—LwF [29], EWC [16], Modified EWC [45], GPM [30], and SGP [31]—on CIFAR100-LT (Table 1), CIFAR10-LT (Table 2), and ImageNet-LT (Table 3). Class sample sizes decrease exponentially, as detailed in Appendix D. Results confirm that CL methods are effective for LTR, aligning with our theorems. While not outperforming specialized LTR methods, CL shows significant improvement over baselines. The superior performance of some of the LTR methods is attributed to tailored design and the potential inexactness of the strong convexity assumption. Existing LTR solutions like BBN learn Head and Tail sequentially to prevent performance loss on the Head. Our results indicate that CL methods are more effective for LTR tasks, suggesting CL could offer more robust solutions. Techniques like RIDE perform well in LTR due to ensemble learning, and CL also can be integrated into such setups. We then utilize the Caltech256 dataset [48] to evaluate the performance of CL on a naturally skewed dataset. The results are presented in Table 4. We observe that CL outperforms the state-of-the-art on this dataset, demonstrating the strong potential of using CL in dealing with long-tailed real-world datasets.

Table 4: The performance of CL compared with SOTA models.

Method	Backbone Inc.V4	Res.101
L^2 – *FE* [46]	84.1%	85.3%
L^2 [46]	85.8%	87.2%
L^2 – *SP* [46]	85.3%	87.2%
DELTA [46]	86.8%	88.7%
TransTailor [47]	-	87.3%
Continual Learning	87.56%	88.9%

4 Conclusion and Future Work

We advanced a CL-based approach for LTR, grounded in the following three theorems that provide insights into optimization dynamics of models in LTR scenarios: 1) an upper bound on weight distances when trained on the Head versus the entire dataset, 2) an extension to multiple subsets, and 3) a proof that CL yields lower loss in LTR scenarios. Our empirical validation on benchmarks like MNIST-LT, CIFAR100-LT, CIFAR10-LT, and ImageNet-LT, as well as real-world data via Caltech256, corroborates our theoretical framework. Future work will delve into non-convex loss landscapes and refine CL methods for LTR, aiming for robust solutions in imbalanced settings.

Acknowledgments

We would like to thank Geotab Inc., the City of Kingston, and NSERC for their support of this work.

References

[1] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[2] William J Reed, "The pareto, zipf and other power laws," *Economics letters*, vol. 74, no. 1, pp. 15–19, 2001.

[3] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng, "Deep long-tailed learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[4] Yu Fu, Liuyu Xiang, Yumna Zahid, Guiguang Ding, Tao Mei, Qiang Shen, and Jungong Han, "Long-tailed visual recognition with deep models: A methodological survey and evaluation," *Neurocomputing*, 2022.

[5] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong, "Long-tailed recognition via weight balancing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6897–6907.

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[7] Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

[8] Chengjian Feng, Yujie Zhong, and Weilin Huang, "Exploring classification equilibrium in long-tailed object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3417–3426.

[9] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu, "Large-scale long-tailed recognition in an open world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2537–2546.

[10] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert, "Learning to model the tail," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] Yaoyao Zhong, Weihong Deng, Mei Wang, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang, "Unequal-training for deep face recognition with long-tailed noisy data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7812–7821.

[12] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong, "Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7610–7619.

[13] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[14] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9268–9277.

[15] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang, "Long-tailed classification by keeping the good and removing the bad momentum causal effect," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1513–1524, 2020.

[16] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al., "Overcoming catastrophic forgetting in neural networks," *National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[17] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9268–9277.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[19] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *International Conference on Learning Representations*, 2019.

[20] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[21] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9719–9728.

[22] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar, "Long-tail learning via logit adjustment," in *International Conference on Learning Representations*, 2020.

[23] Yuzhe Yang and Zhi Xu, "Rethinking the value of labels for improving class-imbalanced learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19290–19301, 2020.

[24] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang, "Long-tailed classification by keeping the good and removing the bad momentum causal effect," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1513–1524, 2020.

[25] Tianhao Li, Limin Wang, and Gangshan Wu, "Self supervision to distillation for long-tailed visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 630–639.

[26] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei, "Distilling virtual examples for long-tailed recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 235–244.

[27] Dvir Samuel and Gal Chechik, "Distributional robustness loss for long-tail learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9495–9504.

[28] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia, "Parametric contrastive learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 715–724.

[29] Zhizhong Li and Derek Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[30] Gobinda Saha, Isha Garg, and Kaushik Roy, "Gradient projection memory for continual learning," in *International Conference on Learning Representations*, 2020.

[31] Gobinda Saha and Kaushik Roy, "Continual learning with scaled gradient projection," *arXiv preprint arXiv:2302.01386*, 2023.

[32] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[33] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4109–4118.

[34] Natalia Díaz-Rodríguez, Vincenzo Lomonaco, David Filliat, and Davide Maltoni, "Don't forget, there is more than forgetting: new metrics for continual learning," 2018.

[35] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[36] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang, "Contrastive learning based hybrid networks for long-tailed image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 943–952.

[37] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia, "Improving calibration for long-tailed recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16489–16498.

[38] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang, "Balanced contrastive learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6908–6917.

[39] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[40] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 181–196.

[41] Si Liu, Risheek Garrepalli, Thomas Dietterich, Alan Fern, and Dan Hendrycks, "Open category detection with pac guarantees," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3169–3178.

[42] Liuyu Xiang, Guiguang Ding, and Jungong Han, "Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 247–263.

[43] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin, "Seesaw loss for long-tailed instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9695–9704.

[44] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun, "Distribution alignment: A unified framework for long-tail visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2361–2370.

[45] Mahdiyar Molahasani, Ali Etemad, and Michael Greenspan, "Continual learning for out-of-distribution pedestrian detection," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 2685–2689.

[46] Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, and Jun Huan, "Delta: Deep learning transfer using feature map with attention for convolutional networks," in *International Conference on Learning Representations*, 2018.

[47] Bingyan Liu, Yifeng Cai, Yao Guo, and Xiangqun Chen, "Transtailor: Pruning the pre-trained model for improved transfer learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 8627–8634.

[48] Gregory Griffin, Alex Holub, and Pietro Perona, "Caltech-256 object category dataset," *Technical report*, 2007.

[49] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1*. Springer, 2005, pp. 878–887.

[50] Chris Drummond, Robert C Holte, et al., "C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," in *Workshop on learning from Imbalanced Datasets II*, 2003, vol. 11, pp. 1–8.

[51] Li Shen, Zhouchen Lin, and Qingming Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. Springer, 2016, pp. 467–482.

[52] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang, "Deep imbalanced learning for face recognition and attribute prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2781–2794, 2019.

[53] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[54] Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao, "Striking the right balance with uncertainty," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 103–112.

[55] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker, "Feature transfer learning for face recognition with under-represented data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5704–5713.

[56] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu, "Long-tailed recognition by routing diverse distribution-aware experts," in *International Conference on Learning Representations*, 2020.

[57] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng, "Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34077–34090, 2022.

[58] Han-Jia Ye, De-Chuan Zhan, and Wei-Lun Chao, "Procrustean training for imbalanced deep learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 92–102.

[59] Emanuele Francazi, Marco Baity-Jesi, and Aurelien Lucchi, "A theoretical analysis of the learning dynamics under class imbalance," *International Conference on Machine Learning*, 2023.

[60] Syed Shakib Sarwar, Aayush Ankit, and Kaushik Roy, "Incremental learning in deep convolutional neural networks using partial network sharing," *IEEE Access*, vol. 8, pp. 4615–4628, 2019.

[61] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong, "Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3925–3934.

[62] Jaehong Yoon, Saehoon Kim, Eunho Yang, and Sung Ju Hwang, "Scalable and order-robust continual learning with additive parameter decomposition," in *Eighth International Conference on Learning Representations, ICLR 2020*. ICLR, 2020.

[63] Gobinda Saha, Isha Garg, Aayush Ankit, and Kaushik Roy, "Space: Structured compression and sharing of representational space for continual learning," *IEEE Access*, vol. 9, pp. 150480–150494, 2021.

[64] Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li, "Orthogonal gradient descent for continual learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 3762–3773.

[65] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro, "Learning to learn without forgetting by maximizing transfer and minimizing interference," in *International Conference on Learning Representations*, 2018.

[66] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny, "Efficient lifelong learning with a-GEM," in *International Conference on Learning Representations*, 2019.

[67] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang, "Online class-incremental continual learning with adversarial shapley value," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 9630–9638.

[68] Chris Dongjoo Kim, Jinseo Jeong, and Gunhee Kim, "Imbalanced continual learning with partitioning reservoir sampling," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 2020, pp. 411–428.

[69] Xialei Liu, Yu-Song Hu, Xu-Sheng Cao, Andrew D Bagdanov, Ke Li, and Ming-Ming Cheng, "Long-tailed class incremental learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 495–512.

[70] Uri Sherman, Tomer Koren, and Yishay Mansour, "Optimal rates for random order online optimization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2097–2108, 2021.

[71] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[72] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[73] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

Appendix

A Related Work

Long-Tailed Recognition. Real-world datasets often exhibit imbalanced distributions, with some classes appearing more frequently than others. Training a model on such imbalanced data can result in poor performance on the rare classes. LTR addresses this issue by enabling models to perform well on both Head and Tail classes [13]. LTR approaches can be broadly categorized into three primary groups: data distribution re-balancing, class-balanced losses, and transfer learning from Head to Tail [19]. Data distribution re-balancing techniques include over-sampling the Tail [6, 49], under-sampling the Head [50], and class-balanced sampling [51, 40]. Class-balanced loss approaches modify the loss function to treat each sample differently, e.g., including class distribution-based loss [13, 14, 52], focal loss [53], and Bayesian uncertainty [54]. Additionally, transfer learning techniques leverage features learned from the Head to improve learning on the Tail [55, 9]. More recently, [21] discussed the limitations of class re-balancing and proposed the Bilateral-Branch Network (BBN) to improve representation learning. [56]. introduced the RoutIng Diverse Experts (RIDE) model to enhance Long-Tailed Recognition (LTR) by reducing model variance. [57] challenged the assumption that test set distribution is always uniform, introducing test-agnostic long-tailed recognition. They used self-supervised learning to facilitate universal feature learning, improving performance on test sets with unknown distribution. Although numerous prior works have addressed LTR, few provide a mathematical analysis of the training process using imbalanced data [58, 59]. These works demonstrate that the Head is learned more quickly than the Tail, primarily focusing on the training dynamics. In contrast, our theoretical analysis studies the convergence point of training within the LTR framework. As mentioned earlier, some of the LTR solutions fall into the category of sequential learning, where head and tail are learned sequentially. Unlike these works, our work delves into the theoretical foundations of why sequential learning is particularly well-suited for LTR, identifying the key factors that influence the success of these methods. By establishing a mathematical framework, we present a novel perspective on the applicability of sequential learning to LTR, a depth of exploration not found in prior works. We also introduce CL as a comprehensive solution to the LTR problem for the first time, drawing on broad principles rather than specific techniques.

Continual Learning. CL addresses the challenge of adapting a deep learning model to new tasks (e.g., new classes or distributions) while maintaining performance on the previously learned tasks. The main challenge to address by CL methods is the mitigation of catastrophic forgetting, i.e., forgetting the previous tasks as the new tasks are learned. CL methods are typically grouped into three categories: expansion-based, regularization-based, and memory-based approaches. Expansion-based CL methods utilize a distinct subset of parameters for learning each task [60, 61, 62]. Regularization-based techniques penalize significant changes in crucial network parameters (relative to previous tasks) by incorporating a regularization term in the loss function [30, 63, 64, 16, 29]. Memory-based approaches employ a replay memory to store a limited number of samples from previous tasks, which are then used in future training to minimize forgetting [65, 66, 67]. Few works attempt to solve both problems of CL and LTR simultaneously in the long-tailed class incremental learning setup. First, [68] have proposed a novel replay method called Partitioning Reservoir Sampling (PRS), which dedicates a sufficient amount of memory to tail classes in order to avoid catastrophic forgetting in minority classes. Class incremental learning is also addressed in a more challenging setup where the new tasks are not uniformly distributed [69]. In this case, the new tasks are LTR, which makes CL more challenging. They considered two setups: Ordered and Shuffled, where the number of samples in each new task is less than in previous tasks, and when the size of classes is completely random, respectively. More recently, gradient surgery has been employed for addressing CL where the gradient from the new task is projected to the orthogonal direction of the previously learned tasks to ensure learning the new task does not impact the previous task [5, 31]. These methods achieve state-of-the-art performance on the CL benchmarks. Note that none of the above works attempts to employ CL as an alternative solution for LT.

B Proofs

B.1 Proof of Theorem 2.2

Proof. The model is trained on the entire dataset $\mathcal{D}$ by minimizing the loss function $\mathcal{L}$:

$$\mathcal{L}(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\left(\sum_{i=1}^{|\mathcal{D}_H|}\ell(\mathcal{D}_H^i) + \sum_{i=1}^{|\mathcal{D}_T|}\ell(\mathcal{D}_T^i)\right), \tag{1}$$

where $\ell(\mathcal{D}^i)$ is the loss of each individual sample. By substituting $\mathcal{L}(\mathcal{D}_H) = \frac{1}{|\mathcal{D}H|}\sum_{i=1}^{|\mathcal{D}_H|}\ell(\mathcal{D}_H^i)$ and $\mathcal{L}(\mathcal{D}_T) = \frac{1}{|\mathcal{D}T|}\sum_{i=1}^{|\mathcal{D}_T|}\ell(\mathcal{D}_T^i)$:

$$\mathcal{L}(\mathcal{D}) = \frac{|\mathcal{D}_H|}{|\mathcal{D}|}\mathcal{L}(\mathcal{D}_H) + \frac{|\mathcal{D}_T|}{|\mathcal{D}|}\mathcal{L}(\mathcal{D}_T). \tag{2}$$

We define $\gamma = \frac{IF}{1+IF}$, which falls within the range of $[0.5, 1)$. We can rewrite Eq. 2 as:

$$\mathcal{L}(\mathcal{D}) = \gamma\mathcal{L}(\mathcal{D}_H) + (1-\gamma)\mathcal{L}(\mathcal{D}_T). \tag{3}$$

Since $IF \gg 0$ in LTR, we can conclude that the value of γ approaches one. Consequently, $\mathcal{L}(\mathcal{D})$ approaches $\mathcal{L}(\mathcal{D}_H)$ for all θ values. Let δ be defined as the maximum difference of the losses:

$$|\mathcal{L}(\mathcal{D}) - \mathcal{L}(\mathcal{D}_H)| \leq \delta. \tag{4}$$

From Eq. 3, it follows that $\lim_{IF\gg 0}\delta = 0$.

One of the most effective losses for the LTR problem is the regularized cross-entropy loss. This loss is the cross-entropy with an additional regularization term that prevents weights from growing excessively:

$$\mathcal{L}(\mathcal{D},\theta) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(P(f(\theta,x_i))\right) + \frac{\mu}{2}\|\theta\|^2, (x_i,y_i) \in \mathcal{D}. \tag{5}$$

This loss improves generalizability by reducing overfitting and achieves state-of-the-art performance when dealing with LTR scenarios [5]. Moreover, as our model is logistic regression, this loss is strongly convex since $\nabla^2\mathcal{L}(\beta,\theta) \geq \mu$. From the definition of strong convexity [70], it therefore follows that:

$$\mathcal{L}(x_1) \geq \mathcal{L}(x_2) + \nabla\mathcal{L}(x_2)^T(x_1 - x_2) + \frac{\mu_{\mathcal{L}}}{2}\|x_1 - x_2\|^2, \tag{6}$$

where $\mu_{\mathcal{L}}$ is the strong convexity parameter. Now, we are introducing Lemma B.1:

Lemma B.1. *If $|f(x) - g(x)| \leq \delta$ and both $f(x)$ and $g(x)$ are strongly convex then:*

$$\|x_g - x_f\|^2 \leq \frac{4\delta}{\mu_f + \mu_g}, \tag{7}$$

where x_g and x_f are $\arg\min f(x)$ and $\arg\min g(x)$, respectively. The proof of this lemma is presented in Appendix B.3.

Applying Lemma B.1 to Eqs. 4 and 6 yields:

$$\|\theta^* - \theta_H^*\|^2 \leq \frac{4\delta}{\mu_H + \mu}, \tag{8}$$

where θ^* and θ_H^* are $\arg\min\mathcal{L}$ and $\arg\min\mathcal{L}_H$, respectively. □

B.2 Remark B.2 and its proof

Remark B.2. *Under a more relaxed assumption, where $\mathcal{L}(\mathcal{D},\theta)$ is strictly (but not strongly) convex, the upper bound can be calculated using Lemma B.3.*

Lemma B.3. *If $|f(x) - g(x)| \leq \delta$ and both $f(x)$ and $g(x)$ are strictly convex then:*

$$\|x_g - x_f\|^2 \leq \frac{4\delta}{\lambda_f + \lambda_g}, \tag{9}$$

where x_g and x_f are $\arg\min f(x)$ and $\arg\min g(x)$, and λ_f and λ_g are the minimum eigenvalues of the hessian matrices of $f(x)$ and $g(x)$, respectively. The full proof is provided in Appendix B.7.

Using lemma B.3, the upper bound of $\|\theta^* - \theta^*_H\|^2$ is expressed as $\frac{4\delta}{\lambda_f + \lambda_g}$. To ensure that this upper bound is limited and approaches zero when $\delta \to 0$, the minimum eigenvalues of the Hessians of both loss functions should have lower bounds, which is again another definition of strong convexity.

B.3 Proof of Lemma B.1

Proof. Since $f(x)$ is strongly convex:

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu_f}{2}\|x_2 - x_1\|^2. \tag{10}$$

Accordingly if $x_2 = x_g = \arg\min g(x)$ and $x_1 = x_f = \arg\min f(x)$, then:

$$f(x_g) - f(x_f) \geq \nabla f(x_f)^T (x_g - x_f) + \frac{\mu_f}{2}\|x_g - x_f\|^2. \tag{11}$$

Since x_f is the minimizer of f, $\nabla f(x_f) = 0$. Therefore:

$$f(x_g) - f(x_f) \geq \frac{\mu_f}{2}\|x_g - x_f\|^2. \tag{12}$$

Similarly, considering $g(x)$, with $x_1 = x_g$, and $x_2 = x_f$, we can derive Equation 10 as follows:

$$g(x_f) - g(x_g) \geq \frac{\mu_g}{2}\|x_f - x_g\|^2. \tag{13}$$

By adding and rearranging Eqs. 12 and 13, we will have:

$$(g(x_f) - f(x_f)) + (f(x_g) - g(x_g)) \geq \frac{(\mu_f + \mu_g)}{2}\|x_g - x_f\|^2. \tag{14}$$

Using $|f(x) - g(x)| \leq \delta$, we can maximize $(g(x_f) - f(x_f))$ and $(f(x_g) - g(x_g))$ to obtain:

$$2\delta \geq \frac{\mu_f + \mu_g}{2}\|x_g - x_f\|^2. \tag{15}$$

Hence:

$$\|x_g - x_f\|^2 \leq \frac{4\delta}{\mu_f + \mu_g}, \tag{16}$$

which completes the proof. □

B.4 Proof of Lemma B.3

Proof. Using the second-order Taylor series expansion for multivariate functions, we can approximate $f(x_g)$ and $g(x_f)$ as follows:

$$f(x_g) \simeq f(x_f) + \nabla f(x_f)(x_g - x_f) + \frac{1}{2}(x_g - x_f)^\top H_f(x_f)(x_g - x_f), \tag{17}$$

$$g(x_f) \simeq g(x_g) + \nabla g(x_g)(x_f - x_g) + \frac{1}{2}(x_f - x_g)^\top H_g(x_g)(x_f - x_g), \tag{18}$$

where $H_f(x_f)$ and $H_g(x_g)$ are the Hessian matrices of f and g evaluated at x_f and x_g, respectively. Since $\nabla f(x_f) = \nabla g(x_g) = 0$, by adding Eq. 17 and Eq. 18 together, we obtain:

$$f(x_g) - g(x_g) + g(x_f) - f(x_f) \simeq \frac{1}{2}(x_g - x_f)^\top H_f(x_f)(x_g - x_f) + \frac{1}{2}(x_f - x_g)^\top H_g(x_g)(x_f - x_g), \tag{19}$$

Using $|f(x) - g(x)| \leq \delta$, we can maximize $(g(x_f) - f(x_f))$ and $(f(x_g) - g(x_g))$:

$$2\delta \geq \frac{1}{2}(x_g - x_f)^\top H_f(x_f)(x_g - x_f) + \frac{1}{2}(x_f - x_g)^\top H_g(x_g)(x_f - x_g), \tag{20}$$

Let λ_f and λ_g be the minimum eigenvalues of $H_f(x_f)$ and $H_g(x_g)$, respectively. By properties of the minimum eigenvalues, we can say:

$$(x_g - x_f)^\top H_f(x_f)(x_g - x_f) \geq \lambda_f \|x_g - x_f\|^2, \tag{21}$$

$$(x_f - x_g)^\top H_g(x_g)(x_f - x_g) \geq \lambda_g \|x_f - x_g\|^2. \tag{22}$$

Using Eqs. 21 and 22, we can rewrite Eq. 20:

$$2\delta \geq \frac{1}{2}\lambda_f \|x_g - x_f\|^2 + \frac{1}{2}\lambda_g \|x_f - x_g\|^2. \tag{23}$$

Therefore:

$$\|x_f - x_g\|^2 \leq \frac{4\delta}{\lambda_f + \lambda_g}, \tag{24}$$

which completes the proof. □

B.5 Proof of Theorem 2.3

Proof. Let $\mathcal{D}$ be a dataset divided into a sequence of partitions $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n$ such that the imbalance factor between any two consecutive partitions $\mathcal{D}_i$ and $\mathcal{D}_{i+1}$ is significantly large, i.e., $\frac{|\mathcal{D}_i|}{|\mathcal{D}_{i+1}|} \gg 1$.

Consider a random subset of $\mathcal{D}$ sorted from largest to smallest denoted as $\mathcal{D}_a, \mathcal{D}_b, \mathcal{D}_c, \ldots$ (where $|\mathcal{D}_a| \gg |\mathcal{D}_b| \gg |\mathcal{D}_c|$).

From Theorem 2.2, we know that if the imbalance factor between two partitions is significantly large, $\frac{|\mathcal{D}_1|}{|\mathcal{D}_2|} \gg 1$, then the distance between the optimal parameters when trained on $\mathcal{D}_1$ and $\mathcal{D}_1 \cup \mathcal{D}_2$ is bounded by ζ, i.e., $\|\theta^*_{\mathcal{D}_1} - \theta^*_{\mathcal{D}1\cup\mathcal{D}2}\|^2 \leq \zeta$ where ζ is computed using Eq. 9 in the manuscript.

Applying this Theorem to $\mathcal{D}_a$ and $\mathcal{D}_b$, we have:

$$\|\theta^*_{\mathcal{D}_a} - \theta^*_{\mathcal{D}_a\cup\mathcal{D}_b}\|^2 \leq \zeta_1$$

Next, considering the combination of $\mathcal{D}_a \cup \mathcal{D}_b$ and $\mathcal{D}_c$, given that $\frac{|\mathcal{D}_a\cup\mathcal{D}_b|}{|\mathcal{D}_c|} \gg 1$, we deduce:

$$\|\theta^*_{\mathcal{D}_a\cup\mathcal{D}_b} - \theta^*_{\mathcal{D}_a\cup\mathcal{D}_b\cup\mathcal{D}_c}\|^2 \leq \zeta_2$$

Given that the weights reside in a metric space, and the distances are Euclidean, the triangle inequality applies. Combining the above inequalities, we therefore get:

$$\|\theta^*_{\mathcal{D}_a} - \theta^*_{\mathcal{D}_a\cup\mathcal{D}_b\cup\mathcal{D}_c}\|^2 \leq (\sqrt{\zeta_1} + \sqrt{\zeta_2})^2$$

Extending this argument for all partitions, we can conclude:

$$\|\theta^*_{\mathcal{D}_a} - \theta^*_{\sum\mathcal{D}_i}\|^2 \leq (\sum_{i=1}^{m}\sqrt{\zeta_i})^2$$

where m is the number of subsets selected randomly. □

B.6 Proof of Theorem 2.4

Proof. Define the updated weight vector after one iteration over the Tail using EWC loss as:

$$\theta^{i+1}_{\text{EWC}} = \theta^i - \eta\nabla\mathcal{L}_{\text{EWC}}(\mathcal{D}_T, \theta^i) \tag{25}$$

Similarly, for $\mathcal{L}$:

$$\theta^{i+1}_{\mathcal{L}} = \theta^i - \eta\nabla\mathcal{L}(\mathcal{D}_T, \theta^i) \tag{26}$$

From the Taylor series expansion, we can estimate the $\mathcal{L}$ of the model with $\theta^{i+1}_{\text{EWC}}$ over $\mathcal{D}$:

$$\mathcal{L}(\mathcal{D}, \theta^{i+1}_{\text{EWC}}) \simeq \mathcal{L}(\mathcal{D}, \theta^i) - \eta \nabla \mathcal{L}_{\text{EWC}}(\mathcal{D}_T, \theta^i) \nabla \mathcal{L}(\mathcal{D}, \theta^i) \tag{27}$$

Similarly, for the $\mathcal{L}$ of the model with $\theta^{i+1}_{\mathcal{L}}$ over $\mathcal{D}$:

$$\mathcal{L}(\mathcal{D}, \theta^{i+1}_{\mathcal{L}}) \simeq \mathcal{L}(\mathcal{D}, \theta^i) - \eta \nabla \mathcal{L}(\mathcal{D}_T, \theta^i) \nabla \mathcal{L}(\mathcal{D}, \theta^i) \tag{28}$$

Subtracting Eq. 28 from 27, we derive:

$$\mathcal{L}(\mathcal{D}, \theta^{i+1}_{\text{EWC}}) - \mathcal{L}(\mathcal{D}, \theta^{i+1}_{\mathcal{L}}) \simeq \eta \nabla \mathcal{L}(\mathcal{D}, \theta^i)(\nabla \mathcal{L}(\mathcal{D}_T, \theta^i) - \nabla \mathcal{L}_{\text{EWC}}(\mathcal{D}_T, \theta^i)) \tag{29}$$

Elastic Weight Consolidation (EWC) loss is expressed as:

$$\mathcal{L}_{\text{EWC}}(\theta^i) = \mathcal{L}(\theta^i) + \frac{\lambda}{2} \sum_i^{|\theta|} F_i (\theta^i - \theta^*)^2 \tag{30}$$

Thus, we can compute $\nabla \mathcal{L}_{\text{EWC}}(\mathcal{D}_T, \theta^i)$ as:

$$\nabla \mathcal{L}_{\text{EWC}}(\mathcal{D}_T, \theta^i) = \nabla \mathcal{L}(\mathcal{D}_T, \theta^i) + \lambda \text{diag}(F)(\theta^i - \theta^*) \tag{31}$$

Substituting Eq. 31 into Eq. 29, we obtain:

$$\mathcal{L}(\mathcal{D}, \theta^{i+1}_{\text{EWC}}) - \mathcal{L}(\mathcal{D}, \theta^{i+1}_{\mathcal{L}}) = -\eta \lambda \text{diag}(F) \nabla \mathcal{L}(\mathcal{D}, \theta^i)^T (\theta^i - \theta^*) \tag{32}$$

To determine the sign of $\eta \lambda \text{diag}(F) \nabla \mathcal{L}(\mathcal{D}, \theta^i)^T (\theta^i - \theta^*)$, we must investigate the sign of each factor. The values of η and λ are positive by construction. To determine the sign of $\nabla \mathcal{L}(\mathcal{D}, \theta^i)^T (\theta^i - \theta^*)$, based on the strong convexity of $\mathcal{L}$ with respect to θ^i and θ^*, we have:

$$\mathcal{L}(\mathcal{D}, \theta^*) \geq \mathcal{L}(\mathcal{D}, \theta^i) + \nabla \mathcal{L}(\mathcal{D}, \theta^i)^T (\theta^* - \theta^i) + \frac{\mu_{\mathcal{L}}}{2} |\theta^i - \theta^*|^2. \tag{33}$$

Rearranging, we obtain:

$$\nabla \mathcal{L}(\mathcal{D}, \theta^i)^T (\theta^* - \theta^i) \leq \mathcal{L}(\mathcal{D}, \theta^*) - \mathcal{L}(\mathcal{D}, \theta^i) - \frac{\mu_{\mathcal{L}}}{2} \|\theta^i - \theta^*\|^2. \tag{34}$$

Since θ^* minimizes $\mathcal{L}$, the term $\mathcal{L}(\mathcal{D}, \theta^*) - \mathcal{L}(\mathcal{D}, \theta^i)$ is always negative. Moreover, $-\frac{\mu_{\mathcal{L}}}{2} \|\theta^i - \theta^*\|^2$ is also always negative, leading to:

$$\nabla \mathcal{L}(\mathcal{D}, \theta^i)^T (\theta^* - \theta^i) < 0. \tag{35}$$

Consequently, $\nabla \mathcal{L}(\mathcal{D}, \theta^i)^T (\theta^i - \theta^*)$ is positive definite.

Finally, the $\text{diag}(F)$ term is determined to be positive valued, according to the following Lemma B.4. Thus we have derived that the sign of $\eta \lambda \text{diag}(F) \nabla \mathcal{L}(\mathcal{D}, \theta^i)^T (\theta^i - \theta^*)$ is positive, which from Eq. 32 we can conclude:

$$\mathcal{L}(\mathcal{D}, \theta^{i+1}_{EWC}) - \mathcal{L}(\mathcal{D}, \theta^{i+1}_{\mathcal{L}}) < 0, \tag{36}$$

which completes the proof. □

Lemma B.4. *Let a logistic regression model be characterized by parameters θ and trained using regularized cross-entropy loss. Then, the diagonal values of its Fisher information matrix (diag(F)) are strictly positive.*

B.7 Proof of Lemma B.4

Proof. The Fisher information matrix is the estimated value of the Hessian of the log-likelihood:

$$F = \mathbb{E}\left[\nabla^2(-\log \mathcal{L}(\theta))\right] \tag{37}$$

In logistic regression, we model the probability of a binary outcome y given input $\mathbf{x}$ as:

$$P(y = 1|\mathbf{x}; \theta) = \frac{1}{1 + e^{-\theta^T \mathbf{x}}} \tag{38}$$

where θ is the vector of model parameters. For a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N\}$, the negative log-likelihood is:

$$-\log \mathcal{L}(\theta) = \sum_{i=1}^{N}\left[-y_i \log\left(\frac{1}{1 + e^{-\theta^T \mathbf{x}_i}}\right) - (1 - y_i)\log\left(1 - \frac{1}{1 + e^{-\theta^T \mathbf{x}_i}}\right)\right] \tag{39}$$

So the Hessian of the negative log-likelihood is:

$$\nabla^2(-\log \mathcal{L}(\theta)) = \begin{bmatrix} \frac{\partial^2(-\log \mathcal{L})}{\partial \theta_1^2} & \cdots & \frac{\partial^2(-\log \mathcal{L})}{\partial \theta_1 \partial \theta_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2(-\log \mathcal{L})}{\partial \theta_d \partial \theta_1} & \cdots & \frac{\partial^2(-\log \mathcal{L})}{\partial \theta_d^2} \end{bmatrix} \tag{40}$$

As a result:

$$\nabla^2(-\log \mathcal{L}(\theta)) = \nabla^2 L(\theta) \tag{41}$$

where d is the dimensionality of θ. Now since the model is logistic regression and loss is regularized cross-entropy, from Eq. 6, we have:

$$\mathcal{L}(x_1) \geq \mathcal{L}(x_2) + \nabla\mathcal{L}(x_2)^T(x_1 - x_2) + \frac{\mu_{\mathcal{L}}}{2}\|x_1 - x_2\|^2, \tag{42}$$

Which is the condition of strong convexity. As a result:

$$\nabla^2 \mathcal{L} \geq \mu_{\mathcal{L}} \boldsymbol{I} \tag{43}$$

From Eq.41 and Eq. 43:

$$\nabla^2(-\log \mathcal{L}(\theta)) = \nabla^2 L(\theta) \geq \mu I \tag{44}$$

Hence:

$$\mathbb{E}\left[\nabla^2(-\log \mathcal{L}(\theta))\right] \geq \mu I \tag{45}$$

consequently:

$$\text{diag}(F) > \text{diag}(D), \quad \text{where } D_{ii} > 0, \quad \text{for all } i \tag{46}$$

which completes the proof. □

C Experimental Setup

Datasets. First, we use the **MNIST-LT** [71] toy dataset with different *IF* values and strong convexity parameters to study the behavior of the upper bound and compliance with our theorem. Next, to evaluate the performance of CL in addressing LTR, we employ three widely used LTR datasets: **CIFAR100-LT**, **CIFAR10-LT** [13], and **ImageNet-LT** [9]. These datasets represent long-tailed versions of the original CIFAR100, CIFAR10, and ImageNet datasets, maintaining the same number of classes while decreasing the number of samples per class using an exponential function. Finally, to highlight the benefits of using CL for LTR, we carry out additional experiments using the naturally skewed **Caltech256** dataset [48].

Implementation Details. We adhere to the experiment setup described in [5, 4]. Following the experimental setup of [5, 4], We use ResNet-32 [72] and ResNeXt-50 [73] for CIFAR and ImageNet benchmarks, respectively. The LTR methods selected for comparison are state-of-the-art solutions in the area. All training was conducted using an NVIDIA RTX 3090 GPU with 24GB VRAM. The details of the implementation specifics are provided in Appendix B.

Evaluation. For the LTR datasets (MNIST-LT, CIFAR100-LT, CIFAR10-LT, ImageNet-LT), we first train the model on the long-tailed imbalanced training set and then evaluate it on the balanced test set,

following the evaluation protocol of [5]. For Caltech256, we use the entire training set for training and assess the model's performance on the entire test set, retaining its original distribution. All reported values represent classification accuracy percentages. All our experiments were conducted utilizing the PyTorch framework[1]. The specifics of each algorithm's implementation are summarized in Table B1. The parameters for each algorithm such as Learning Rate (LR), Optimizer, Momentum, LR Scheduler, CL Weight, and number of Epochs are detailed. The algorithms considered include Learning without Forgetting (LwF), Elastic Weight Consolidation (EWC), a modified version of EWC, Gradient Projection Memory (GPM), and Scaled Gradient Projection (SGP).

Table A1: Implementation Details of the Considered Algorithms for LTR benchmark.

Algorithm	LR	Opt.	Momentum	LR Scheduler	CL Loss Weight	Epochs
LwF	0.001	SGD	0.9	-	0.01	5
EWC	0.01	SGD	0.9	-	10	90
Modified EWC	0.01	SGD	0.9	-	1000	90
GPM	0.001	SGD	0	Cosine Anneal LR	-	100
SGP	0.001	SGD	0	Cosine Anneal LR	-	150

D Datasets

Fig. C1 illustrates the distribution of samples among different classes and the division of the dataset into the Head and Tail sections. In the case of CIFAR100-LT with $IF = 100$, the initial partition is configured such that 5% of the samples fall within the Tail and 95% in the Head section (Classes 60 to 100 are classified as Tail). For comparison purposes, the rest of the datasets follow a similar partition threshold where 60% of the classes are assigned to the Head section.

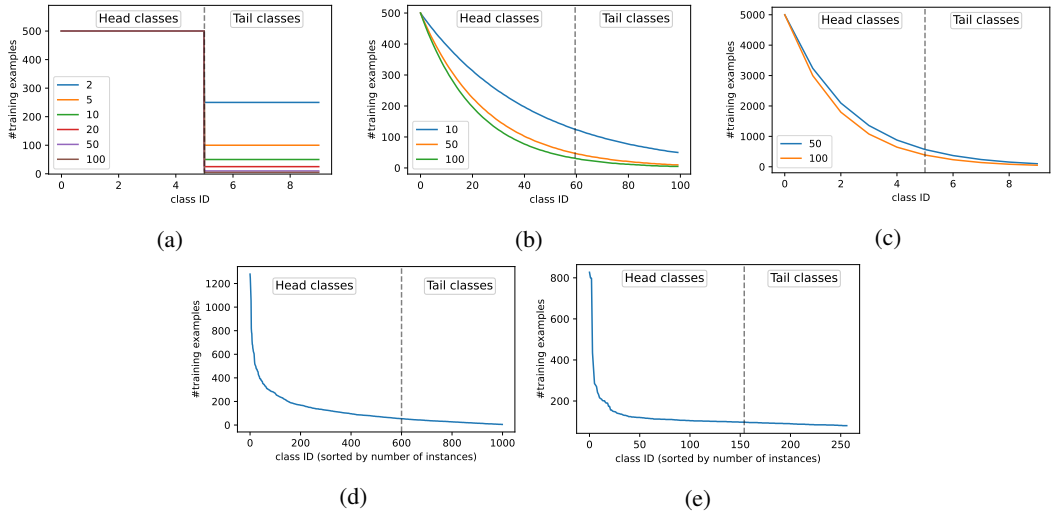


Figure A1: Class cardinality of (a) MNIST-LT, (b) CIAFR100-LT, (c) CIFAR10-LT, (d) ImageNet-LT and (e) Caltech256

E The Impact of Imbalance Factor and strong Convexity Parameter

To investigate the distance between the acquired sets of weights by training on $\mathcal{D}$ or $\mathcal{D}_H$ ($\|\theta^* - \theta^*_H\|$), we first train a logistic regression model on MNIST-LT with varying IF and μ values. Then we calculate the Euclidean distance between the two sets of weights, as illustrated in Fig. 4. As expected from Eq. 8, increasing either the IF or strong convexity (μ) results in a reduced distance, indicating

[1]The code for the algorithms was obtained and modified from various open-source repositories:
https://github.com/ngailapdi/LWF
https://github.com/shivamsaboo17/Overcoming-Catastrophic-forgetting-in-Neural-Networks
https://github.com/sahagobinda/GPM
https://github.com/sahagobinda/SGP

that the weights of the model trained using $\mathcal{D}$ approach the weights when it is solely trained using $\mathcal{D}_H$.

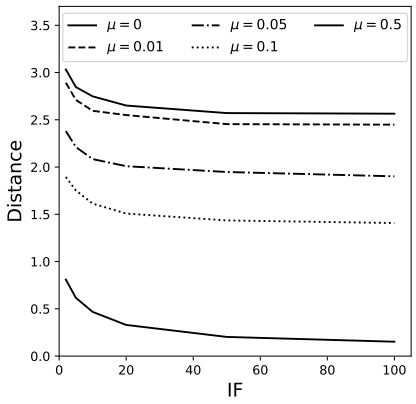


Figure 4: The distance between θ^* and θ^*_H in different *IF* and μ.

F CL Performance Analysis

Here, let's discuss three key concepts in the context of CL: catastrophic forgetting, backward transfer, and forward transfer [34]. As mentioned earlier, catastrophic forgetting occurs when the performance of a class declines after retraining. Despite the use of CL methods, which are designed to mitigate this forgetting, a certain degree of forgetting is still inevitable. Forward transfer is the improvement in performance on a new task after employing CL, which is the central aim of retraining in CL. Finally, backward transfer is a beneficial side-effect where retraining on new samples can actually enhance the model's performance on the previous tasks. Now, let's discuss Fig. 5, which presents the difference in per-class accuracy of the best CL method (SGP) versus the baseline network. The analysis is based on CIFAR100-LT with an *IF* of 100. The figure is divided into three regions corresponding to the scenarios discussed above: catastrophic forgetting (bottom), backward transfer (top-left), and forward transfer (top-right). The bottom region in the figure represents classes that undergo catastrophic forgetting, while the top-right region represents the Tail samples (with a class index larger than 60), which demonstrate improved performance, or forward transfer. We observe that using SGP as a CL solution for LTR results in very effective improvements in the per-class accuracy of the Tail (forward transfer). Interestingly, despite the absence of Head data in the retraining process, 42 out of 60 Head classes see some level of improvement after the model is exposed to the Tail samples (backward transfer). This result emphasizes the remarkable potential of CL methods in enhancing the performance on both new and previous tasks.

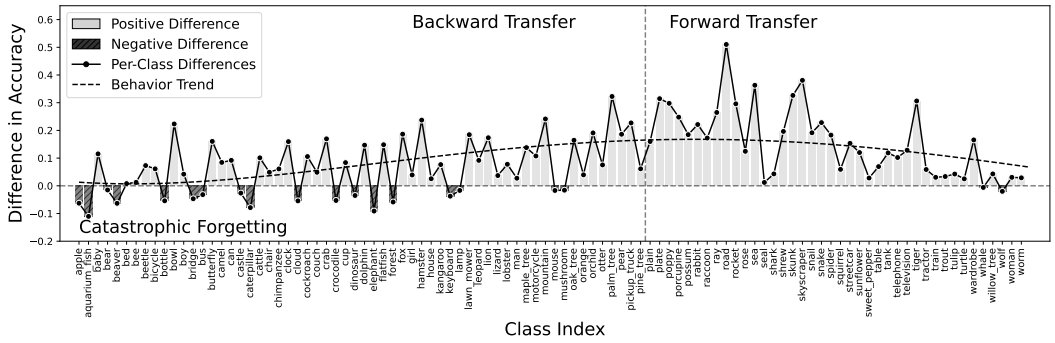
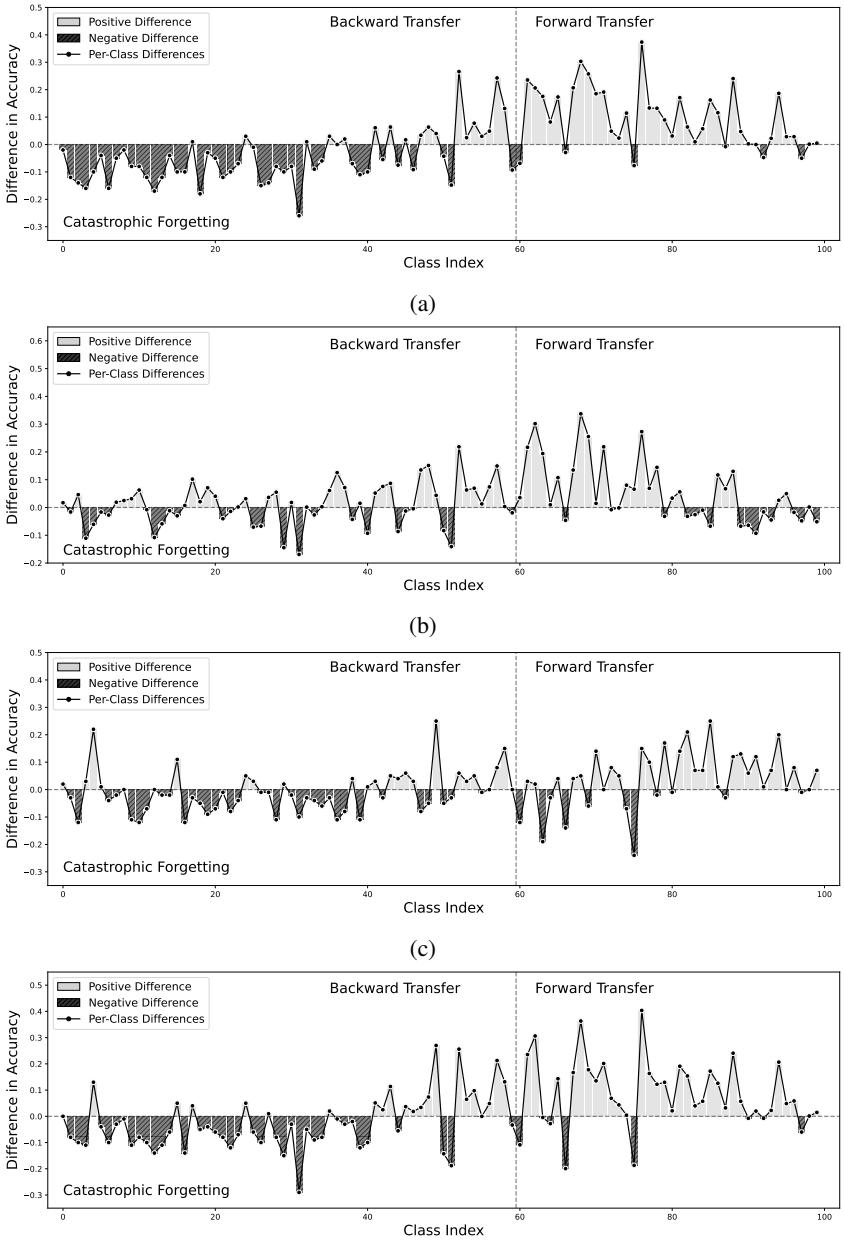


Figure 5: The difference in per-class accuracy of SGP and the baseline model.

Here, rather than employing the baseline for computing per-class accuracy differences, we compare the CL method, GPM (which follows the same trend but slightly worse performance than SGP), with an LTR model, WD, that exhibits similar overall accuracy. The outcomes are depicted in Fig. A2 (a). In this figure, the red bars denote classes where WD outperforms GPM, whereas the bluebars indicate the classes where GPM excels. We observe that GPM performs generally better on the Tail,

whereas WD outperforms in Head. On average, WD's accuracy on Head classes is 4.5% higher, while GPM achieves a 9.5% higher accuracy on Tail samples. Here, we analyze the difference in per-class accuracy of GPM , Modified EWC (which exhibits similar but slightly better performance than EWC), and LwF with respect to each other, and present the results in Figs. A2 (b, c, and d). Among these three CL methods, GPM demonstrates the best results on the Tail, particularly in classes 60 to 80. LwF performs better when data is extremely limited (classes 90 to 100). The best method for Head classes is Modified EWC (outperforming GPM in 40 out of 60 Head classes), as a result of both minimizing instances of catastrophic forgetting and promoting backward transfer. These comparisons highlight that each CL method exhibits distinct behaviors when applied to the LTR problem.



(a)

(b)

(c)

(d)

Figure A2: The difference in per-class accuracy of (a) GPM and WD, (b) GPM and LwF, (c) LwF and Modified EWC, and (d) GPM and Modified EWC.

G Run Time analysis

The inference runtime is identical between CL-based methods and LTR solutions, due to the same architecture in both types of methods and the fact that CL does not affect inference. When CL is used to address LTR, the data is divided into head and tail sets. At each step of the training, only one partition of data is involved. Since the architecture is consistent among all LTR approaches within a particular benchmark, the runtime is determined by the amount of data fed to the model. So, dividing the learning into multiple steps and using CL does not impact the total runtime, nor does it increase the training time.

H Weight imbalance

An interesting phenomenon observed when training models on highly imbalanced data is the presence of artificially large weights in neurons corresponding to the Head classes [5]. The LTR solution, WD, addresses this problem by penalizing weight growth using weight decay. One way to assess the network's ability to handle LTR is by analyzing the bias in per-class weight norms. To this end, we present the per-class weight norms of the Baseline, WD, and SGP models in Fig. 7.

The figure reveals a significant imbalance in the weight norms of the Baseline model, which is naively trained on the imbalanced dataset. In contrast, the WD and SGP models exhibit more uniform weight norms across different classes. Interestingly, although SGP starts with the heavily imbalanced weights of the Baseline model, it converges towards a more uniform weight distribution without any explicit penalty on weight growth. Unlike many other CL methods that restrict the plasticity of crucial weights, GPM only constrains the direction of the weight update in the weight space, enabling the model to converge to a more balanced weight distribution. This further demonstrates the effectiveness of CL in addressing LTR.

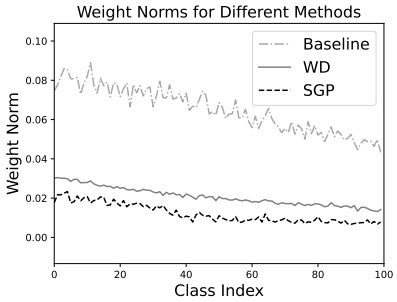


Figure 7: Per-class weight norms of the baseline, SGP, and WD.

I Limitations

Strong convexity is a key assumption in our theorem, which determines an upper bound for the distance between the weights of a learner trained on the full dataset and the weights of the same learner trained solely on the Head. This assumption offers a solid theoretical foundation for our method, showcasing the feasibility of using CL techniques to address the LTR problem. However, as many deep learning models in practice employ non-convex loss functions that potentially limit the theorem's applicability to specific cases, it is crucial to highlight that our experimental results are not strictly dependent on the strong convexity condition. In fact, our method exhibits impressive performance even under more relaxed conditions, indicating its robustness and adaptability.

J Broader Impact

Dealing with imbalanced data is of paramount importance in ensuring fairness and reducing bias in AI applications, particularly in cases where the underrepresented classes correspond to minority groups. The long-tailed distribution of real-world data poses a significant challenge in achieving equitable performance for both common and rare cases. This paper's proposed algorithm, which addresses the LTR problem through the lens of CL, holds great potential in mitigating the adverse effects of class imbalance on model performance. By effectively learning from both the Head and the Tail, the proposed method can enhance the performance on underrepresented classes, leading to more fair and accurate AI models across various domains.